知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

一种结合字面与上下文相似性的招聘网页技能词语规范化方法

◎孙瑜[1] 姜金德[2]
[1] 金陵中学河西分校 南京 210019
[2] 南京晓庄学院商学院 南京 211171

摘要：[目的 / 意义] 针对招聘网页文本存在许多英文技能词语拼写错误的问题，提出一种招聘网页技能词语规范化方法。[方法 / 过程] 结合字面相似性和上下文相似性，度量技能词语的相似度，形成相似技能词语网络，从而对招聘网页文本中的技能词语进行规范化。[结果 / 结论] 从国内主流招聘网站前程无忧获取一周计算机类岗位求职信息，使用提出的方法进行招聘网页英文技能词语规范化。实验结果表明，提出的方法能够自动、准确、快速地规范招聘网页文本中的技能词语。

关键词：招聘网页 技能 词语规范化
分类号：G202

引用格式：孙瑜 , 姜金德 . 一种结合字面与上下文相似性的招聘网页技能词语规范化方法 [J/OL]. 知识管理论坛 , 2018, 3(6): 325-334[引用日期]. http://www.kmf.ac.cn/p/151/.

近几年来，随着我国高等教育的迅猛发展和招生规模的日益扩大，大学生找工作难、企业招人难已经成为社会关注的热点。在某种程度上，我国高校人才培养与社会需求间的不匹配，造成了这种双重困境。特别是在信息时代中，企业对人才的需求变化迅速，与之相矛盾的是高校人才培养周期长，专业课程设置滞后，导致学生的培养脱离实际需要。因此，在高速发展的信息时代中，快速、准确地洞察企业对所招岗位技能需求显得格外重要。随着互联网的普及，网络招聘成为企业招聘的主流方式。招聘网页中常含有企业对所招岗位技能需求的具体描述，反映了当前就业市场对人才的技能需求。因此，通过分析招聘网页信息，了解整个社会对某领域人才技能需求是一种有效的实现途径。由于招聘网页为非结构化的文本，需要进行一系列的自然语言处理操作，从而获取相关的结构化的技能信息。然而，不同于传统的经过严格编辑和修订的文本，网络招聘文本书写通常不规范，特别在一些领域中，技能通常为一些英文词语，存在许多错误拼写，如将“Oracle”错拼为“Orace”、将“Linux”错拼为“Liunx”等。招聘网页文本技能书写的不规范对基于传统规范文本的自然语言处理方法产

作者简介：孙瑜 (ORCID：0000-0001-8275-8824)，学生，E-mail: weifanglai@sina.com；姜金德 (ORCID：0000-0002-5504-7493)，教授，博士。

收稿日期：2018-09-17 发表日期：2018-12-07 本文责任编辑：刘远颖

生了干扰。因此，在对招聘网页文本进行技能需求分析之前，将招聘网页文本中拼写不规范的英文技能词语转换为规范形式显得尤为重要。

近年来，已经有一些研究尝试利用网络招聘信息分析企业招聘岗位对技能的需求 [1-6]。但是，这些研究通常采用手工方式进行技能词语规范，这不能适应招聘网页更新快速、数据量大的特点。目前，对招聘网页中技能词语进行自动规范的研究还较少。文献 [7] 针对招聘文本，提出用词向量聚类的方法进行技能词语规范化处理，然而，该方法没有考虑拼写错误的技能词语通常具有相似的字面形式，并且词向量模型不能很好地为低频词语产生准确的词向量，从而影响技能词语规范化的效果。

通过仔细观察招聘网页文本中的技能词语，可以发现错拼的词语通常具有相似的字面形式，并且具有相似的上下文技能词语术语，因此，笔者提出结合字面和上下文相似性的方法度量技能词语的相似度，形成相似技能词语网络，从而对招聘网页文本中的技能词语进行规范化。实验表明，笔者提出的方法能够自动、准确、快速地规范招聘网页文本中的技能词语。

1 相关研究

词语规范化（lexical normalization）是将多个词语归纳成一个等价类，是众多自然语言预处理的一个重要步骤。例如机器翻译、命名实体抽取、信息检索等研究，它们处理的数据都是经过规范化后的“干净”语料，从而降低模型的复杂度。词语规范化是语料预处理的一个关键步骤，一直以来都备受研究者关注，尤其是随着近年来社交媒体上的文本呈爆炸式增长，社交媒体文本词语规范化成为研究的热点。

早期文本规范化工作大多使用噪声信道模型。文献 [8] 首先将噪声信道模型应用于文本规范化任务，提出一种基于字符串编辑的噪声信道模型，该模型对子串转换的概率建模，可提高文本规范化的效果；文献 [9] 通过扩展噪声信道模型中错误模型（将词之间的语音相似性加入错误模型），通过学习规则来预测每一个字符的发音，并且预测依赖于词中的相邻其他字符。但这种模型为有监督模型，需要大量标注语料对模型进行训练。

以噪声信道模型为基础，目前的词语规范化方法可分为拼写修正法、序列标注法和机器翻译法三大类。拼写修正法假设单词变成非标准词的过程是相互独立的，则文本规范化问题可以简化为单词拼写修正问题。序列标注法将词语规范化任务看作一个序列标注问题进行求解。首先针对文本中的每个单词生成候选的若干个规范化单词，然后采用维特比算法基于语言模型进行求解，得到联合概率最大的单词序列作为规范化结果。通常所采用的序列模型有隐马尔科夫模型 [10] 和条件随机场 [11]。机器翻译法借助词对齐概念，对非标准词—标准词关系中的一对多、多对一和多对多映射进行建模 [12-13]。其中，序列标注和机器翻译方法是有监督的方法，需要大量标注数据训练模型，训练数据需要耗费大量人力进行手工标注。因此，利用非监督的拼写修正方法规范词语成为研究的热点。

拼写修正方法主要包括基于词形相似性和基于上下文相似性两类。基于词形相似性的方法中最具有代表性的是通过计算单词的编辑距离来表示单词相似性，而在社交文本中，非标准词形很可能和标准形式大相径庭，文献 [14] 提出针对社交文本的词语相似性模型，使用近音拼写、单词裁剪等变化形式度量单词相似性。词语的上下文相似性则指不同单词出现在相似上下文中的概率。目前，通常使用神经网络训练出来的词向量 [7, 15-17]。特别地，文献 [7] 提出利用 word2vec 词向量表示技能词语以及上下文，以进行招聘文本技能词语规范化。

目前的拼写修正方法大多针对社交网络文本，相较于社交网络词语的多种形式，招聘网页文本技能词语通常具有字面相似性；而上下文中使用词向量方法并不适用，因为词向量更适合准确地表示高频词语，低频词语的词向量

知识管理论坛，2018(6):325-334
DOI：10.13266/j.issn.2095-5472.2018.031

并不准确，从而影响了上下文词语相似性计算。笔者提出的方法是针对招聘网页文本的无监督的拼写修正方法，不需要使用标注数据，适应招聘网页更新快速、数据量大的特点。

2 结合字面和上下文相似性的技能词语规范化方法

招聘网页中错拼的词语通常具有相似的字面形式，如将“Oracle”错拼为“Orace”等。但是，如果仅仅使用字面相似性也可能造成一些非错拼的词语被误认为是错拼词语，如“Radware”与“Hardware”虽具有相似的字面形式，但是为两个不同的词语，“Radware”为一家领先的智能化解决方案供应商，致力于确保快速、可靠、安全地交付网络或基于网络的应用程序，而“Hardware”表示计算机系统的组成硬件。通过观察，可以发现错拼的词语通常具有相似的上下文词语，而非错拼的相似字面形式词语通常具有不同的上下文技能词语。例如，“Oracle”与“Orace”通常都与“数据库”“SQL”等词一起出现；而词“Radware”通常和“WebLogic”“Bea”“Server”等词同时出现，“Hardware”则和“显示器”“主板”“CPU”“内存”等词一起出现。

因此，针对中文招聘网页文本中存在许多英文技能拼写错误的问题，笔者提出一种结合字面和上下文相似性的技能词语规范化方法，方法总流程如图 1 所示。方法主要分为预处理、计算技能词语对相似性和生成技能词语相似网络等 3 个步骤。首先对获取的招聘网页文本进行相关预处理工作，然后计算技能词语对的字面相似性和上下文相似性，形成技能词语相似性度量，根据词语相似性度量形成相似技能词语网络，以进行技能词语规范化。

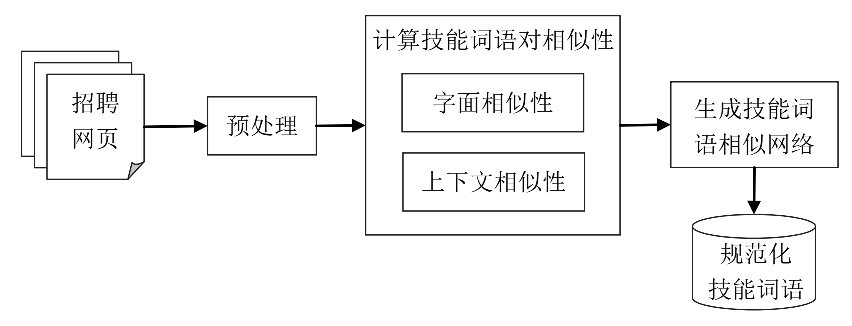


图 1 结合字面和上下文相似性的技能词语规范化方法总流程

2.1 预处理

由于招聘网页文本是非结构化的网页结构，而且除了包含技能等所需信息之外，还包括其他大量噪音信息，如广告、图片动画、与主题无关的超级链接、脚本语言以及各类标签。因此，需要针对招聘网页文本结构，借助网页文本分析工具解析包解析网络文本，提取出招聘网页中与招聘分析信息有关的文本结构。然后，对获取的相关文本内容进行去重、词性标注、英文大小写转化等操作。由于本文旨在对招聘网页文本中的英文技能词语进行规范化，因此，预处理阶段将过滤掉中文词语，仅保留英文技能词语。图 2 为对一个招聘网页文本进行预处理示例，保留了任职要求中的所有英文词语，一个岗位招聘网页文本形成一个对应的岗位技能词语文本。

2.2 计算技能词语对相似性

为了规范技能词语，需要计算技能词语之间的相似度，以判断两个词语为同一技能的可能性。笔者从技能词语字面和上下文两方面衡量技能词语的相似性，当两个词语字面越相似，上下文越相似，则越可能是同一技能词语。

```
<span class="sp4"><emclass="i4"></em>04-21 发 布</span><divclass="clear"></div><divclass="clear"></div></div></div></div><divclass="tBorderTop_box"><h2><span class="bname">职位信息</span></h2><divclass="bmsgjob_msg inbox"><p>任职要求：</p><p>1、1 年以上 Java 开发经验；对软件工程和相关标准有良好的认识；具有较强的面向对象思维；精通设计模式；</p><p>2、熟悉 Spring、MyBatis 等主流 J2EE 技术；熟练使用 Oracle 数据库，并有一定的 SQL 优化经验；</p><p>3、熟悉 Javascript、JQuery、BootStrap、CSS 等技术；熟悉 Linux 操作系统；熟悉 Tomcat 应用服务器；</p><p>4、有 Spark、Hadoop 开发经验者优先</p><p>5、能够承受压力、基础扎实、思路清晰，有独立解决问题的能力、良好的沟通表达能力，有责任心，具有良好的团队合作意识。</p><p>五险一金：享受齐全的社会保险，包括养老、医疗、失业、工伤、生育、以及住房公积金。
```

```
java spring mybatis
j2ee oracle
sqljavascriptjquery
bootstrap csslinus
tomcat spark hadoop
```

图 2 招聘网页文本预处理示例

2.2.1 字面相似性

最具有代表性的基于字面相似性的方法是通过计算词语的编辑距离 (Edit Distance, ED)[18] 来表示单词的相似性。编辑距离指使一个候选技能术语变为另一个候选技能术语而进行的插入、删除、替换等操作的最少次数。编辑距离不仅考虑了两个候选技能术语之间相同字符的数目，还考虑了它们之间位置关系，通常，编辑距离越小，说明两个候选技能术语越相似。如，“oracle”与“orace”的编辑距离为 1。由于编辑距离没有考虑候选技能术语本身长度，因此笔者将两个候选技能术语的长度融入编辑距离，形成归一化编辑距离 (Normalized Edit Distance, NED)，其定义如下：

$$\mathrm{NED}(w_i,w_j)=\frac{ED(w_i,w_j)}{|w_i|+|w_j|} \qquad 公式 (1)$$

在公式（1）中，ED (w_i, w_j) 表示技能词语 w_i 与 w_j 之间的编辑距离。由 NED 定义可知，NED 越小，则两个技能词语术语越相似，当 w_i 和 w_j 完全相同时，其 NED 为 0。“oracle”与“orace”的 NED 值为 $\frac{1}{6+5}=0.09$。

基于归一化编辑距离，词语 w_i 与词语 w_j 之间的字面相似性 strSim 定义如下：

$$strSim(w_i,w_j)=\frac{1}{NED(w_i,w_j)}, w_i \neq w_j \qquad 公式 (2)$$

由公式 (2) 可知，当两个词语归一化距离越小，则字面相似性越大，表明两个词语越可能是同一词语，为了避免分母为 0，不考虑 w_i 和 w_j 完全相同的情况。

2.2.2 上下文相似性

相似技能词语通常有相似的上下文技能词语。因此，可以利用技能词语的上下文判断技能词语的相似度。具体地，给定词语 w_i，D_i 为包含技能词语 w_i 的所有岗位技能词语文本集，即 $D_i=\{D_{i1},...,D_{im},...,D_{in}\}$，$w_i \in D_{im}$。笔者定义两个上下文相似性 conSetSim 和 conFreSim，其中 conSetSim 的定义如下：

$$conSetSim(w_i,w_j)=\frac{|S_i \cap S_j|}{|S_i \cup S_j|} \qquad 公式 (3)$$

在公式（3）中，S_i 为 w_i 的上下文技能词语集，$S_i=\{w \mid w \in D_i \wedge w \neq w_i\}$；同样地，$S_j$ 表示 w_j 的上下文技能词语集。由公式（3）可知，当技能词语 w_i 与技能词语 w_j 的上下文技能词语集中相

知识管理论坛，2018(6):325-334
DOI：10.13266/j.issn.2095-5472.2018.031

同的技能词语越多的时候，则 w_i 与 w_j 越相似，越可能是表示相同概念的技能词语。

conFreSim 的定义如下：

$$conFreSim(w_i,w_j)=\frac{\sum_{w\in S_i\cap S_j} n_i^w\times n_j^w}{\sqrt{\sum_{w\in S_i}(n_i^w)^2}\times\sqrt{\sum_{w\in S_j}(n_j^w)^2}} \quad \text{公式 (4)}$$

在公式（4）中，n_i^w表示词语 $\boldsymbol{w}$ 在词语 $\boldsymbol{w_i}$ 的岗位上下文词语文本集 $\boldsymbol{D_i}$ 中出现的次数；类似地，n_j^w 表示词语 $\boldsymbol{w}$ 在词语 $\boldsymbol{w_j}$ 的岗位上下文词语文本 $\boldsymbol{D_j}$ 中出现的次数。由公式（4）可知，conFreSim 与 conSetSim 的相同之处都是通过上下文词语集来度量词语相似度，但是 conFreSim 还考虑了上下文技能词语集中每个技能词语出现的次数。

综合考虑字面相似性和上下文相似性两个指标，形成最终的指标 strSim-conSetSim 和 strSim-conFreSim 两个相似性指标，它们的定义如下：

$$strSim-conSetSim(w_i,w_j)=strSim(w_i,w_j)\times conSetSim(w_i,w_j) \quad \text{公式 (5)}$$

$$strSim-conFreSim(w_i,w_j)=strSim(w_i,w_j)\times conFreSim(w_i,w_j) \quad \text{公式 (6)}$$

由公式（5）和（6）可知，当 w_i 和 w_j 的字面相似性越大，上下文相似性越大时，指标 strSim-conSetSim 和 strSim-conFreSim 越大，表明 w_i 和 w_j 越可能是相同的技能词语。

2.3 生成相似技能词语网络

根据技能词语对的相似性，可以生成相似技能词语网络，以找到所有表示相同技能概念的技能词语。相似技能词语网络中的每个顶点表示一个技能词语，网络之间的无向边为相似性大于某个预先设定的阈值的技能词语对。根据生成的无向网络，寻找网络中所有的连通网络，即为表示相同概念的技能词语集。使用每个集合中出现频次最高的技能词语术语表示该集合，以进行技能词语规范化操作。图 3 为相似技能词语网络示例。在图 3 的相似技能词语网络中，共有 3 个连通网络，形成 3 个技能词语集合，即 {websphere, webspere, websphare}、{visio、viso、vioso}、{zibbix、zabbix、zabix}。

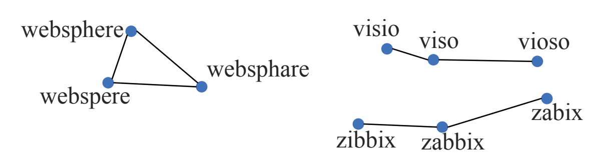


图 3 相似技能词语网络示例

3 实验

3.1 数据集

为了验证笔者提出方法的可行性与有效性，实验抓取国内主流招聘网站前程无忧 (www.51job.com) 招聘网页数据，以规范技能词语。前程无忧是一家网络招聘服务提供商，是中国最具影响力的人才招聘网站之一。按照职能的划分，在前程无忧网站选取“计算机 / 互联网 / 通信 / 电子”职能抓取数据（数据抓取日期：2018-3-19 至 2018-3-26）作为招聘网页集，去除内容重复、全英文、没有写明任职要求的招聘网页，最后共得到 14 678 个相关招聘网页。

3.2 实验步骤与评估方法

实验首先对招聘网页文本进行预处理，包括使用 Beautifulsoap 定位，解析网页内容，获得岗位技能要求文本，使用结巴分词进行词性标注、英文大写转化为小写等工作，最终保留英文技能词语，共得到 7 156 个不同的英文词语。分别计算两两不同技能词语的字面相似性和上下文相似性，最终形成相似度，人工设定阈值为 7，形成相似技能词语网络，找出相似技能词语网络中所有的连通网络，形成技能词语集合，使用各集合中出现频次最高的词语作为规范词语进行规范化。

实验对技能词语对相似性进行人工标注，判断其是否为同一技能概念，采用 P@N 方法评价正确评估的技能词语对，其公式如下：

$$P@N=\frac{\text{\#前N对相似技能词汇对中为同一技能概念}}{\text{\#前}N\text{对相似技能词汇对}}\times 100\% \quad \text{公式 (7)}$$



知识管理论坛
2018 年 第 6 期(总第 18 期)

3.3 结果

3.3.1 技能词语相似性方法评估

实验首先评估第 3.2 节中提出的技能词语相似性方法，为此分为 4 组分别进行比较，分组如表 1 所示：

表 1 相似方法方法

序号	比较方法	计算方法
第 1 组	strSim	公式（2）
	conSetSim	公式（3）
	conFreSim	公式（4）
第 2 组	strSim	公式（2）
	conSetSim	公式（3）
	strSim-conSetSim	公式（5）
第 3 组	strSim	公式（2）
	conFreSim	公式（4）
	strSim-conFreSim	公式（6）
第 4 组	strSim	公式（2）
	strSim-conSetSim	公式（5）
	strSim-conFreSim	公式（6）

第 1 组使用字面相似性和两种上下文相似性分别计算词语对的相似性。其结果如图 4 所示。由图 4 可见，在 3 种相似性计算中，在前 600 对技能词语对中，strSim 方法准确率最高，但是在 600 对技能词语对之后，strSim 方法准确率迅速下降，低于上下文的两类方法 conSetSim 和 ConFreSim。而上下文相似度的两种方法中，conSetSim 好于 conFreSim。strSim 方法利用字面计算词语对的相似性，能够较为准确地得到一些拼写错误的词对，但是也存在一些错误，如，技能词语“spring”和“swing”虽然字面相似，但并非同一个技能概念，是两种不同的计算机技能词语。conSetSim 和 conFreSim 方法利用词语的上下文判断词语间的相似性，有着较为稳定的准确率，其中 conFreSim 考虑上下文技能词语的词频，能够更加精确地刻画技能词语对的相似性，因此准确率好于 conSetSim 方法。

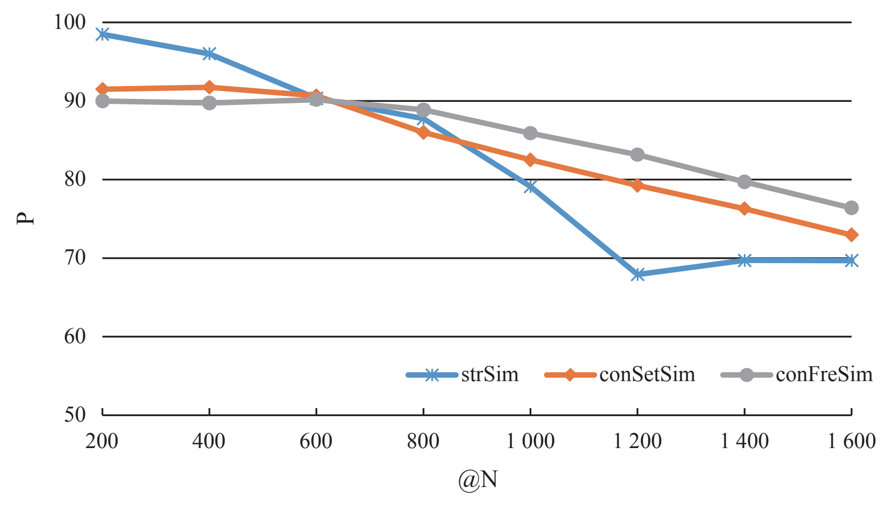


图 4 第 1 组技能词语对相似性方法比较结果

第 2 组评估使用上下文相似度 conSetSim 和字面相似度 strSim 的混合方法 strSim-conSetSim 方法，已评估 strSim-conSetSim 方法是否能够提高准确率（见图 5）。由图 5 可见，结合字面相似性和上下文相似性的 strSim-conSetSim 方法明显提升了准确率，该方法总体也好于单独使用字面相似性 strSim 和单独使用上下文相似性 conSetSim 的方法。这表明，字面相似性从字面上计算词语对的相似性，而上下文相似性从词语上下文计算词语对的相似性，两者相互补充，能够取得更好的结果。

知识管理论坛，2018(6):325–334
DOI：10.13266/j.issn.2095–5472.2018.031

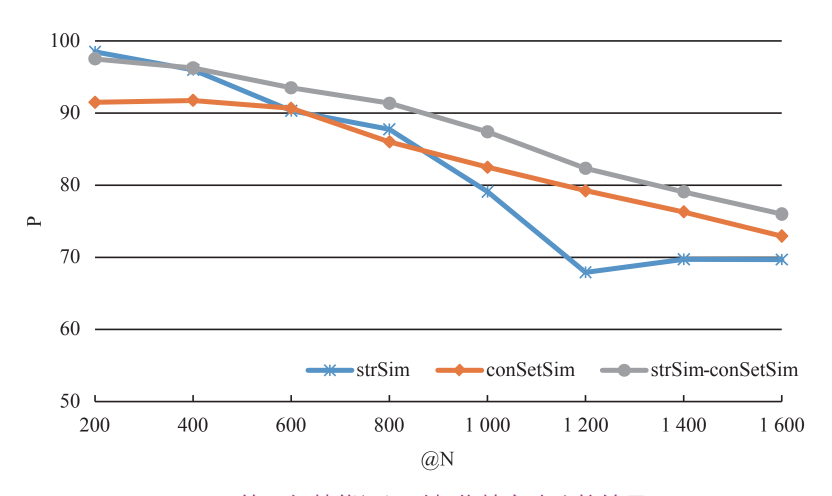


图 5 第 2 组技能词语对相似性方法比较结果

第 3 组方法比较了上下文相似性 conFreSim 和字面相似性 strSim 结合的方法 strSim-conFreSim 对性能的影响（见图 6）。由图 6 可见，结果与第 2 组结果类似，结合字面相似性和上下文相似性的 strSim-conFredSim 方法明显提升了准确率，该方法总体好于单独使用字面相似性和单独使用上下文相似度的方法。这进一步证明结合字符相似性与上下文相似性能够取得更好的结果。

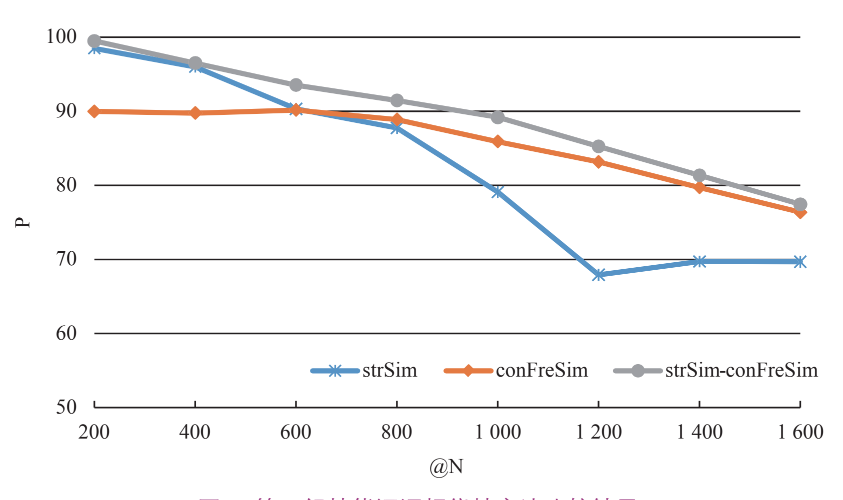


图 6 第 3 组技能词语相似性方法比较结果

第 4 组方法比较了两种字面相似性和上下文相似度结合方法，即 strSim-conSetSim 和 strSim-conFreSim 方法，并使用 strSim 作为基准方法，结果如图 7 所示。由图 7 可见，两种结合方法中，strSim-conFreSim 方法略好于 strSim-SetSim，此结果与第 1 组方法中 conFreSim 方法好于 conFreSim 方法结论一致。相较于 conSetSim 方法，conFreSim 方法考虑上下文技能词语的频次，能够更加精确地刻画词语，因此准确率好于 conSetSim 方法。

3.3.2 与其他方法比较

接着，实验使用第 3.3.1 节中最佳方法 strSim-conFreSim 与文献 [7] 中的方法进行了比较，并使用 strSim 方法作为基准方法。文献 [7] 使用神经网络词向量方法，利用技能词语的上下文计算词语向量，以规范化招聘文本技能词语，笔者将其简称为 word2vecSim 方法。实验结果如图 8 所示。由图 8 可见，笔者提出的方法好

于 word2vecSim 方法，strSim 方法也优于 word-2vecSim 方法。原因主要有 3 个：① word2vec 使用神经网络计算词向量，虽然考虑了上下文，但是 word2vec 只有当技能词语大量出现时，才能获得较好的准确度，而当技能出现较少时，则性能并不好；② word2vec 考虑的上下文不仅仅有技能词语，还有其他非技能词语，这样会影响技能词语的利用，从而不能准确地产生技能词语向量；③ word2vec 没有考虑技能词语对的字面相似性。

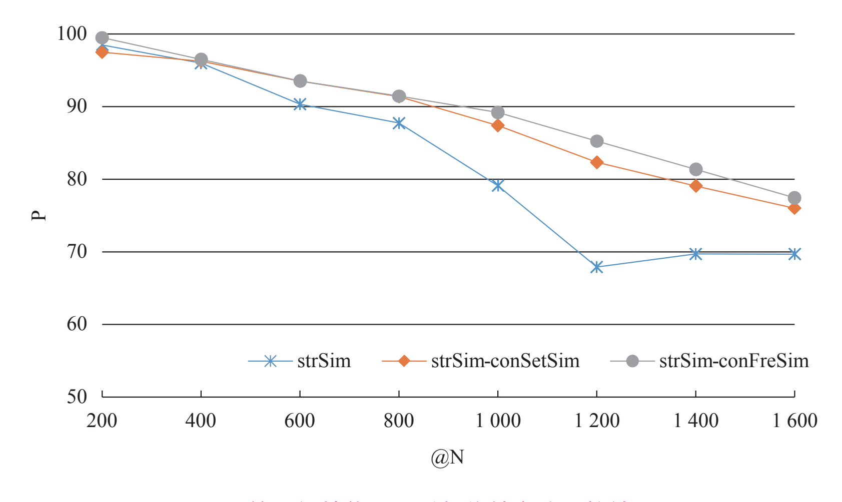


图 7　第 4 组技能词语对相似性方法比较结果

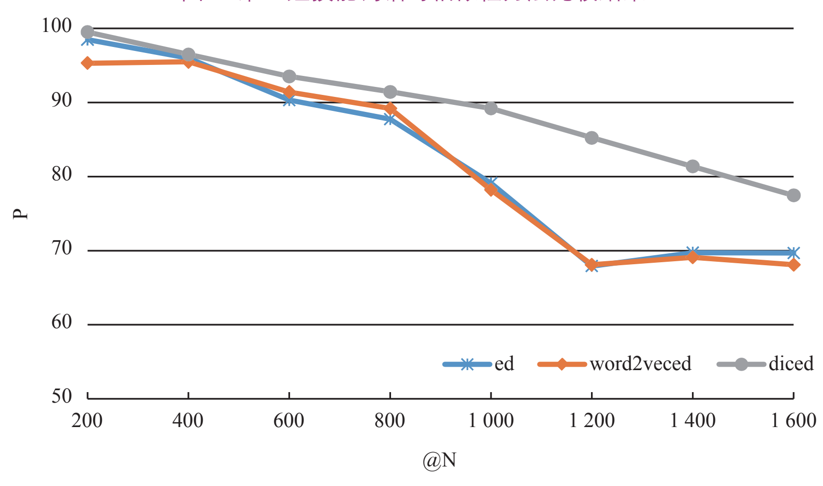


图 8　与其他相似性方法比较结果

3.3.3 实际案例分析

表 2 列出了 3 个实际案例。表 2 中的第 1 对词语对，具有较大的字面相似度，且具有较大的上下文相似度，从而最后的 strSim-confreSim 相似度值较高，可以判断它们为同一技能概念。表 2 中的第 2 对词语对，虽然具有较大的字面相似度，但实际为两个不同的技能词语概念，它们的上下文相似度并不高，最终的 strSim-conFreSim 相似度不高，因此不是同一技能概念。表 2 中的第 3 对词语对为两个相关的技能词语，因此具有较高的上下文相似性，但是两者字面相似性很小，可以判断不是同一技能概念。这表明，通过结合字面相似性和上下文相似性能够获得比单独使用其中一种方法更准确的技能词语相似性。

知识管理论坛，2018(6):325-334
DOI：10.13266/j.issn.2095-5472.2018.031

表 2 实例分析

序号	词	词	strSim	conFreSim	strSim-conFreSim
1	oracle	oralce	6.00	3.15	18.90
2	fireworks	framework	6.33	0.59	3.73
3	mapreduce	hadoop	2.14	2.67	5.71

3.4.4 规范化实例

最后，实验通过相似技能词语网络，找到连通网络，形成若干技能词语集，采用集合中最大词频作为标准化词语，表 3 列出部分规范化实例。由表 3 结果可见，笔者提出的结合字相似性和上下文相似性的方法，能够很好地规范化招聘网页中拼写错误的技能词语。

表 3 词语规范化实例

序号	技能词语集	规范化词	序号	技能词语集	规范化词
1	android andriod anroid andorid andoird androin andiod androd androrid androi andoid	android	4	jquery jquey jqueyr jqurey jqeury juqery juqery jqery jquer	jquery
2	eclipse eclips eclise eclipe elipse	eclipse	5	hibernate hibenate herbinate hibernat hiberate hibernet	hibernate
3	mybatis mybaits mybaties myibatis mybaitis mybiatis mybtis mybastis mbatis mybits	mybatis	6	struts structs stucts struct strut strus strust	struts
4	javascript javascrip javascrpt jscript javasript javasscript javascipt javacript javascripts javascrpit javasprict javascrtip javscript javescript javascritp	javascript	8	oracle oralce oracel orcale orcle orcal oraccle oracl orale oracal orace	oracle

4 总结

网络招聘信息中常含有企业对所招岗位技能需求的具体描述，反映了当前就业市场对人才的技能需求。因此，通过分析网络招聘信息，了解整个社会对某领域人才技能需求是一种有效的实现途径。然而，不同于传统的经过严格编辑和修订的文本，网络招聘文本书写通常不规范，特别在一些领域中，有关技能的描述通常为一些英文，产生许多错误拼写。网络招聘文本技能词语书写不规范的特点，基于传统规范文本的自然语言处理方法再使用中会受到干扰。因此，在网络招聘文本进行技能需求分析之前，将网络招聘文本中拼写错误的英文技能词语转换为规范形式显得尤为重要。笔者提出结合字面相似性和上下文相似性的方法度量技能词语的相似度，根据技能词语的相似度，形成相似技能词语网络，从而规范化招聘网页文本中的技能词语。从国内主流招聘网站前程无忧获取一周计算机类岗位求职信息，使用提出的方法进行招聘网页英文技能词语规范化。实验结果表明，笔者提出的方法能够自动、准确、快速地规范化网络招聘文本中的技能词语。从而进行招聘岗位技能需求分析和知识发现，化解就业知识供需不对称问题，帮助高等院校和大学生合理有效地利用网上就业信息资源，帮助高校专业管理者快速洞察企业对专业人才的技能需求，为其制定符合企业需求的专业人才培养方案提供情报决策支持。

参考文献：

[1] WOWCKO I. Skills and vacancy analysis with data mining techniques[J]. Informatics, 2015, 2(4):31-49.

[2] KIM J Y, LEE C K. An empirical analysis of requirements for data scientists using online job postings[J]. International journal of software engineering and its



application, 2016, 10(4): 161-172.

[3] 夏火松，潘筱听．基于 Python 挖掘的大数据学术研究与人才需求的关系研究 [J]. 信息资源管理学报，2017, 7(1): 4-12.

[4] 詹川．基于文本挖掘的专业人才技能需求分析——以电子商务专业为例 [J]. 图书馆论坛，2017, 5(1): 116-123.

[5] 夏立新，楚林，王忠义，等．基于网络文本挖掘的就业知识需求关系构建 [J]. 图书情报知识，2016, 169(1):94-100.

[6] 刘睿伦，叶文豪，高瑞卿，等．基于大数据岗位需求的文本聚类研究 [J]. 数据分析与知识发现，2017, 12(12): 32-40.

[7] LUO Q, ZHAO M, JAVED F, et al. Macau: large-scale skill sense disambiguation in the online recruitment domain[C]// IEEE international conference on big data. Piscataway: IEEE, 2015:1324-1329.

[8] BRILL E, MOORE R C. An improved error model for noisy channel spelling correction[C]// Meeting of the Association for Computational Linguistics. Piscataway: IEEE, 2000:286-293.

[9] TOUTANOVA K, MOORE R C. Pronunciation modeling for improved spelling correction[C]// Proceedings of annual meeting of the Association for Computational Linguistics. Stroudsburg :Association for Computational Linguistics, 2002:144-151.

[10] CHOUDHURY M, SARAF R, JAIN V, et al. Investigation and modeling of the structure of texting language[J]. International journal of document analysis & recognition, 2007, 10(3):157-174.

[11] LIU F, WENG F, WANG B, et al. Insertion, deletion, or substitution? normalizing text messages without pre-categorization nor supervision[J]. 2012, 15(2):71-76.

[12] AW A T, ZHANG M, XIAO J, et al. A phrase-based statistical model for SMS text normalization.[C]// International conference on computational linguistics and meeting of the Association for Computational Linguistics. New York: ACM, 2006: 17-21.

[13] PENNELL D L, LIU Y. A character-level machine translation approach for normalization of sms abbreviations[J]. Natural language processing, 2011,20(2):974-982.

[14] COOK P, STEVENSON S. An unsupervised model for text message normalization[M]. Stroudsburg：Association for computational linguistics, 2009.

[15] SRIDHAR V K R. Unsupervised text normalization using distributed representations of words and phrases[C]// The workshop on vector space modeling for natural language processing. Piscataway: IEEE, 2015:8-16.

[16] 施振辉，沙灜，梁棋，等．基于字词联合的变体词规范化研究 [J]. 计算机系统应用，2017, 26(10):29-35.

[17] 罗延根，李晓，蒋同海，等．基于词向量的维吾尔语词项归一化方法 [J]. 计算机工程，2018(2):220-225.

[18] DAMERAU F J. A technique for computer detection and correction of spelling errors[J]. Communications of the ACM, 1964, 7(3):171-176.

作者贡献说明：
孙　瑜：提出研究思路，实施实验，撰写论文；
姜金德：分析实验数据，修改论文，进行理论指导。

A Skill Vocabulary Normalization Method for Recruitment Webpage Combing Literal and Context Similarity

Sun Yu[1] Jiang Jinde[2]
1. Jinling Middle School Hexi School, Nanjing 210019
2. School of Business, Nanjing Xiaozhuang University, Nanjing 211171

Abstract: [Purpose/significance] This paper proposes a skill vocabulary normalization method for recruitment webpages, it aims to solve the problem that many English skill word spelling errors exist in the recruitment webpages. **[Method/process]** The method combines literal similarity and context similarity to measure the similarity of skill word and form a similar skill word network to normalize the skill words in the recruitment webpages. **[Result/conclusion]** One week's computer recruitment information was obtained from domestic mainstream recruitment website 51job to evaluate the proposed method. The experiment results show that the proposed method can automatically, accurately and quickly normalize the skill vocabulary in the recruitment webpages.

Keywords: recruitment webpage　skill　lexical normalization